Author Names & Affiliations

- Ping Li - Pennsylvania State University
- Jenni Evans - Pennsylvania State University
- Vasant Honavar - Pennsylvania State University
- Guido Cervone - Pennsylvania State University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Psychology, neuroscience, computational modeling, machine learning, earth and environmental science, artificial intelligence

Title of Submission

Cyberinfrastructure Needs to Address Science and Engineering Research Challenges

Abstract (maximum ~200 words).

How should a large public institution such as Penn State take advantage of advanced cyberinfrastructure to tackle challenges that impact our society and the world and to enable important scientific advances? In this response, we identify the emerging interdisciplinary research initiatives that rely on cyberinfrastructure, outline the significant challenges of building a responsive cyberinfrastructure ecosystem, and provide suggestions on the strategies and directions to address the sustainability of cyberinfrastructure for cyber-enabled scientific discoveries in the 21st century.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Today's cutting-edge research tackles critical and complex problems facing our society, utilizing highly interdisciplinary data- and computation-intensive approaches. These approaches rely on big data, and on the cyber-ecosystem that can support model building and analyses of the data. Through cyber-enabled scientific advances, we might be able to:

1) Engineer precise medical treatments of diseases with predictable outcomes;
2) Create personalized models of learning to optimize the learning outcomes of individual students;
3) Control the flow of energy in new types of devices, resulting in dramatic increases in device capabilities without increasing energy needs;
4) Develop strategies for sustainable development;

5) More effectively respond to man-made or natural disasters;

6) Synthesize experimental data into computational models for designing and predicting the outcomes of new experiments;

7) And achieve many other societally impactful and scientifically important goals.

Many challenges stand in the way of these advances. One critical problem is that our ability to collect massive volumes and different varieties of complex data has outstripped our capacity to analyze them. Another is that we increasingly need simulations that can span vast ranges of scale.

In explaining and contextualizing these challenges, we draw upon cases faced by researchers connected to Penn State's Institute for CyberScience (ICS), a university-wide research organization supporting interdisciplinary cyber-enabled science and engineering research. ICS accomplishes its integrative mission both by growing an interdisciplinary community of cyber-enabled researchers and cyber-professionals, and by providing this community with a state-of-the-art high performance cloud (the ICS Advanced CyberInfrastructure, ICS-ACI).

Data Deluge: The Volume of Data Outstrips our Capacity to Analyze It

Advances in instrumentation and methodology in many fields have led to a data-rich, but knowledge-poor, research environment. Researchers now have an unprecedented access to data, but challenges remain in integrating, analyzing, and understanding this data.

In meteorology and Earth Science in general, the advances in our ability to observe and model the Earth system through the use of remote sensing, numerical models and social media has led to the generation of geographically distributed, asynchronous, dynamic, and diverse datasets. The rate at which these data are generated often greatly exceeds our ability to analyze them.

Likewise, new research in the social and behavioral sciences increasingly depends on understanding the social networks that are characterizing a new era of human communication through social media. Social media data can provide critical insights into human behavior and organization, including political view and policy evolution, socio-economic development, population changes, and many more. This data is voluminous and accumulates by the second.

The emerging field of "precision health in context" relies on understanding the individual's psychological, cognitive, and neuroanatomical states through big data models. A wealth of data is available from censuses, diagnoses, tests, biomonitoring devices (such as FitBits), and biomedical imaging, but this data is far quicker to collect than to analyze and use.

The life sciences are being transformed by various advances in method and instrumentation. High-throughput laboratory techniques have yielded vast amounts of genetic information, and have made possible system-wide measurements of the entire transcriptome, large portions of the proteome, interactome, metabolome, and diseasome. Advances in imaging, including the development of cryo-electron microscopes, have made possible the collection of diverse cell images at, which can be combined with measurement of the expression levels of thousands of genes to directly see the responses of individual cells and tissues under different perturbations. These advances, while promising, are also extremely data-intensive.

These are only a few examples of the many fields that face the challenge of data deluge. Advanced analytical tools, algorithms, and computational models are imperative for ingesting and analyzing all of the available data.

Expanding the Scalability of Simulations

A key application of cyberscience techniques involves predictive modeling in a framework of uncertainty. From forecasting weather and natural disasters to disease outbreaks and financial risk, predictions are hampered by incomplete definition of the initial conditions – even in the face of massive advances in observational data streams – and our ability to represent the complete range of scales necessary to minimize uncertainty in predictions.

For instance, an accurate simulation of a hurricane needs to span processes from molecular to planetary scales. Interactions at the atomic level can have significant impacts on a hurricane's growth, as the interactions of water molecules and phase changes of water drive the convective clouds that govern hurricane intensity. But hurricane models must also be able to simulate the large-scale interactions of a hurricane with other weather systems, as these govern the storm's motion and structure.

Many other fields face similar multi-scale modelling challenges. Predicting the dynamics of cancer might require simulations that combine

cellular-scale with organism-scale models. Predicting how new materials and chemical compounds interact might require combining models at different scales of time, from femtoseconds to seconds.

Despite the promise of widely-scalable simulations, modeling frameworks that can handle vastly different scales and which can deliver accurate predictions in uncertain conditions are computationally intensive and demand personnel with a high level of expertise. This increasing level of demand but lack of corresponding supply of human expertise may present significant challenges to many fields that rely on advanced cyberinfrastructure.

Society faces challenges that are fundamental to the health and wellbeing of individuals, communities, nations and the planet itself. Increasingly rapid advances in observing and sensing technologies raise research questions that would have been beyond imagination in earlier eras, but developing the capability to process the data created by these technologies remains difficult. Integrating the data into multi-scale models that accurately predict uncertain events is also a major challenge. Solving these problems will likely lead to major transformation in a variety of fields and have crucial societal impacts.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The research challenges described above point to various cyberinfrastructure needs. While specific cyberinfrastructure requirements differ across fields, some needs are common, including:
1) Standard tools, approaches, and frameworks for handling big data across different disciplines;
2) Development of new tools that incorporate insights from emerging statistical, analysis and machine learning approaches, among others;
3) Solutions for end-to-end management of big data, from ingesting massive amounts of it to archiving the products of data; and
4) Maintaining the security of data, especially as cyber-enabled researchers increasingly rely on clouds.

Standard Tools, Approaches, and Frameworks for Handling Big Data

Many research fields rely on gathering, analyzing, interpreting, and integrating massive volumes of data. Across fields, there are widely different approaches to coordination, orchestration, and execution of these tasks. Many special-purpose tools exist for solving specific data problems, but their proliferation reflects a lack of best practices for developing, standardizing, and executing models. Tools providing a framework for scalable and flexible execution of a large and collective set of tasks for high-level data mining and visualization of big data, applicable to a variety of data types, would be valuable to many fields.

The lack of these standard tools and practices hampers interdisciplinarity. For example, in Earth Science piecemeal solutions have to be replaced by cyberinfrastructure solutions cognizant of the end-to-end application requirements and supporting different science problems that have similar computational characteristics. Several tools have been developed to assist Earth scientists to run their workloads on HPC infrastructures specifically with the requirements of solving a specific problem as the first-order design consideration. These solutions are often tied to a single problem domain and are not applicable to different domains that have similar computational requirements.

Similarly, biology and neuroscience labs are often interested in similar issues related to neurodegenerative disease, each collecting their own data that could be mutually informative (e.g., a biology lab with large amount of neuroanatomical data from mice and rats and a psychology lab with huge human brain imaging data). Analyzing this data is difficult—using conventional methods to image a fruitfly's tens of thousands of neurons might take a decade, while doing the same for the human brain's 86 billion neurons could take 17 million years (Nature 541, Jan. 26 2017). Thus, developing innovative computational data analytics and algorithms (e.g., bioinformatic data-mining) is necessary for making use of the data. These tools would have to handle the different types of data gathered by biologists and neuroscientists; only then would scientists in these fields arrive at cross-cutting insights that will lead to fundamental discoveries of biological/neural processes.

While standard tools are important, future cyberinfrastructure advances must also be flexible enough to meet the needs of different kinds of research projects. Many research projects can benefit from exploiting HPC systems, but given the current design of HPC resources towards single monolithic jobs, not all tasks make optimal use of existing HPC systems. There is a need for robust and scalable

approaches toward large-scale data analytics that can cover the wide-spectrum of application scenarios. We need cyberinfrastructure solutions that support different science problems with varying computational characteristics.

These tools must be accessible, and support for them must be sustainable, especially in today's climate of tight funding. Software tools and data streams would need to be disseminated freely to the research community and frequently updated. The community would need to provide consultation on what hardware and software advances would need to be supported within a standard framework.

End-to-End Data Management Solutions

Cyberinfrastructure could help to relieve the problems caused by the data deluge by streamlining many tasks in the data lifecycle—the stages through which data passes, from initial ingestion to dissemination and archiving. Each step in this lifecycle has its demands. Some raw data sources must be downloaded manually and passed through collection points before being transferred to the storage space where it can be analyzed and manipulated. Sharing datasets with collaborators can often cause delay and replication, and is not always done securely. Disseminating data requires tagging it with metadata—an often complex manual process.

Penn State has begun to design a comprehensive end-to-end data management system that automates many tasks in the data lifecycle. The proposed system would allow for automatic downloading of data from designated sources, as well as automatic data tagging and classification to reflect security/privacy access and other use restrictions (HIPAA, etc.). Ingested data would be routed by the system to a secure shared research space, where it could be accessed and manipulated by teams of collaborators. Once results have been obtained, the system would automate the generation of metadata and push the data to outside portals, where it could be searched for and obtained by other researchers and the public.

This system would help to alleviate strain on intra-institution networks, especially at large institutions like Penn State. Currently, the limited capacity for transmission of terabytes (or more) of data from one site to another, even within the Penn State system, obstructs research progress and discourages collaboration. An integrated data flow framework such as that described here would reduce unnecessary repetitive data transfers and go some way to addressing this problem, but would not completely obviate it.

Related to concerns about the data lifecycle, and to satisfying Federal and other data management plans, is the question of data archiving costs and searchability. Once a grant expires, there is no funding source to support storage of research products created from the work funded by that. Further, codified requirements on metadata and data formats are needed to ensure that data remains searchable in long term. This includes a need to refresh software for these data formats so they do not timeout.

Compliance in the Cloud

Researchers at many institutions are increasingly turning to cloud providers for high-performance computing services. How these cloud environments affect compliance with data security and privacy policies, however, is as yet unclear. Community guidance is needed on institutional responsibilities for data governance when individual researchers use the cloud and when cyber-institutes take research into the cloud.

Addressing these cyberinfrastructure issues—finding standard tools and frameworks that aid collaboration while addressing the demands of different types of computational research; streamlining end-to-end data management; and ensuring cloud compliance—would allow researchers better to take advantage of massive quantities of available data and perform more accurate and complex simulations.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Besides issues that can be addressed through cyberinfrastructure advances, there are other areas of action that NSF ought to consider, including:
1) Support for training to develop skills needed to use cyberinfrastructure effectively
2) A coordinating body to oversee the use of "restricted data"
3) A deliberate framework for coordinating collaborative community endeavors
4) Increasing NSF's agility for planning new cyberinfrastructure-related initiatives

Submission in Response to NSF CI 2030 Request for Information
DATE AND TIME: 2017-04-05 15:00:34
REFERENCE NO: 255

PAGE 5

5) Providing long-term support for cyber-research experts

Researchers and students must be informed of the latest innovative cyberinfrastructure technologies and tools, and they must receive the training that can help them use these tools. This education and training must be made available at all levels—from first degree earners to returning professionals. Stronger coordination between institutes such as ICS (and academic institutions more generally) and industry would help keep researchers on the cutting edge of technology, and allow students to prepare for promising academic or industry careers.

Currently, using restricted data at large academic institutions presents an administrative challenge. At Penn State, a researcher wishing to use data with regulatory and/or contractual restrictions on data use and/or dissemination must work with at least ten different offices, and often more, to remain compliant. Due to the dispersed services of these offices, we do not have a straightforward, routine way to recognize that we are receiving restricted data, categorize the data, implement the technical requirements of the restrictions, and ensure that compliance is met throughout the research project. A coordinating body that could work to streamline the use of restricted data at large institutions would help address this problem.

NSF funds a number of community endeavors for sharing disparate data across related communities (e.g. EarthCube). To make these activities more effective, a deliberate framework for coordinating these activities and sharing information is needed. What progress is being made across the community—and which areas should progress—must be made clear. Otherwise, we end up with unintended outcomes that may satisfy nobody, or a community comprised only of a very small core team (which defeats the original purpose of these endeavors). NSF's community initiatives should include cyberinfrastructure specialists, not just domain experts. While domain experts may have strong programming and data management expertise, their skills are often limited to their research areas; cyber-ecosystem experts can bring new valuable new approaches to research problems.

NSF must remain agile in anticipating emerging needs in existing cyber-heavy disciplinary communities and funding new communities that need a cyberinfrastructure ecosystem. We suggest that, in order to maintain awareness of community needs, NSF gathers 2-3 researchers chosen by each Directorate annually (different people every year) to identify where their field is going in the realm of CI. This feedback would be used to develop agile NSF planning for new collaborative CI-related initiatives, with roadmaps for researchers to develop their research.

Cyber-research experts are becoming an increasingly critical part of today's complex cyber-ecosystem. We need a model for employing these individuals (likely PhD level, many with non-CS PhDs) that does not lead to "holes" in their funding. Institutions can't afford to provide permanent salaries for all research areas, so we need a model that allows for a combination of agency and university funding that is flexible across individual grants. An institution-level grant to support the growth of research in new fields might be a way to provide consistent funding for these experts. This support would allow us to build a more diverse cyber-community, as we will be able to accommodate life transitions (parenthood, gaps in workforce, etc.).

Finally, the cyber-ecosystem should also seriously consider the education of new generations of scientists who are CI-oriented graduate students today. This steady supply of technical expertise to the research fields will be critical for the sustainability of the future of cyberinfrastructure and the advances of frontier sciences and engineering.